International Journal of Electrical and Computer Engineering (IJECE)
Vol. 10, No. 1, February 2020, pp. 353~359
ISSN: 2088-8708, DOI: 10.11591/ijece.v10i1.pp353-359

Performance evaluation of random forest with feature selection methods in prediction of diabetes

Raghavendra S[1], Santosh Kumar J[2]
[1]Department of Computer Science and Engineering, CHRIST Deemed To Be University, India
[2]Department of Computer Science and Engineering, KSSEM, India

Article Info

Article history:

Received Jan 10, 2019
Revised Aug 7, 2019
Accepted Aug 29, 2019

Keywords:

Classification accuracy
Data mining
Feature selection method
Percentage split
Random forest

ABSTRACT

Data mining is nothing but the process of viewing data in different angle and compiling it into appropriate information. Recent improvements in the area of data mining and machine learning have empowered the research in biomedical field to improve the condition of general health care. Since the wrong classification may lead to poor prediction, there is a need to perform the better classification which further improves the prediction rate of the medical datasets. When medical data mining is applied on the medical datasets the important and difficult challenges are the classification and prediction. In this proposed work we evaluate the PIMA Indian Diabtes data set of UCI repository using machine learning algorithm like Random Forest along with feature selection methods such as forward selection and backward elimination based on entropy evaluation method using percentage split as test option. The experiment was conducted using R studio platform and we achieved classification accuracy of 84.1%. From results we can say that Random Forest predicts diabetes better than other techniques with less number of attributes so that one can avoid least important test for identifying diabetes.

Corresponding Author:

Raghavendra S,
Department of Computer Science and Engineering,
CHRIST Deemed To Be University,
Kanmanike, Kumbalgodu, Mysore Road, Bangalore-560074. Contact: +919740857501
Email: raghav.trg@gmail.com

1. INTRODUCTION

Recent improvements in the area of data mining and machine learning have empowered the research in biomedical field to improve the condition of general health care. In many parts of the world the tendency for maintaining long-lasting records consisting of medical data is becoming an accepted practice. In addition to this, the newer medical equipment's and the techniques used in diagnosis, produces composite and huge data. Therefore, to handle these ill-structured biomedical data, intelligent algorithms for data mining and machine learning are required in order to take logical reasoning from the saved raw data, which is considered as medical data mining. Within the medical data, the medical data mining searches for patterns and relationships which can provide useful information for appropriate medical diagnosis [1]. Data mining techniques are applied to different medical domains (health care databases or medical datasets) to improve the medical diagnosis.

To check for any invisible patterns inside the medical datasets, medical data mining is strongly recommended. In medical data mining, the actual tasks (challenges) are the classification and prediction of medical datasets. One of the techniques that is used for the classification and prediction is random forest [2]. Random forest is an ensemble learning method for regression, classification, and other jobs that functions by making an assembly of decision trees at training time and generating the class that is the classification or regression of the distinct trees.

Random forest is a versatile algorithm used for classification suited for the analysis of large datasets. Random forest is popular because random forest classification models have high prediction accuracy and provides information on important variables for classification. Random forest provides two important aspects for data mining i.e. high prediction accuracy and information related to importance of attribute in classification [3].

One of the significant steps in any data mining research based on classification and prediction models is the feature selection. Feature selection is very much important because when we construct a medical data mining model, the medical dataset may generally consists of further information than the actual information needed to construct the model. If we preserve the attribute columns that are not actually needed, then it leads to wastage of memory and more CPU time is needed for the training process and the quality of the explored pattern may be deteriorated by these additional attributes because of the following reason:

a. It is difficult to discover meaningful pattern from data because some attributes may be redundant, and noisy.
b. Foridentifying excellent pattern, the majority of data mining algorithms need larger training dataset butthe data used for training is extremely small in few data mining applications.

Feature selection assist in solving these problems by having too little data of high value rather than too much data of little value.Feature selection has advantages in the classification of data and there is reduction in computational complexity due to reduction in dimension [4].

In the proposed research work we apply data mining technique like random forest with feature selection methods on diabetes dataset and have identified the key attributes for increasing the classification accuracy.

2. LITERATURE REVIEW

From the existing literature we found many different methods are applied on PIMA Indian diabetes dataset. The methods proposed by different researchers and the classification accuracy achieved are explained below:

To exhibit the efficiency of the hybrid classifier based on evolutionary computation on diabetes dataset, a method based on hybrid classifier along with k-nearest neighborwas proposed. Based on the classification accuracy, it was clear that on over 50 runs the hybrid classifier achieved good accuracy of 80% [5].

Instead of using a traditional neuron which produces output for a given input in each iteration, a spiking neuron which gets activated after each T ms with an input is designed. The output can be changed into a particular firing rate furthermore it can perform the data classification depending on a firing rate created from input signal. For a set of cases belonging to one among k classes, every input is connected to input current and the spiking neuron gets excited after T ms, at last the firing amount is calculated for each case. Weights for the spiking neuron are optimized using a gravitational search algorithm. The capability of the projected method is compared with the identical spiking neuron implemented with particle swarm optimization (PSO), cuckoo search algorithm and differential evolutions. The model is implemented on diabetes dataset and the gravitational search algorithm achieved good accuracy of 76.61% [6].

For optimizing the parameter for support vector machine (SVM), an adjusted bat algorithm (ABA) is proposed. The experiments are conducted on the diabetes dataset. The experimental result was compared with the Grid-SVM and other approaches. Based on the result, ABA-SVM is considered as a better classifier than Grid-SVM and compared to other approaches like PSO-SVM, the ABA-SVM achieved better classification accuracy of 77.34% [7].

A model is proposed to handle the problems that can appear when learning from very small data that are already classified. The model depends on Logical Analysis of Data (LAD) and is provided with additional information obtained from the consideration of data statistically. So the new proposed model is called SLAD. The performance of SLAD is compared with LAD, SVM and label propagation algorithm. The experiment was conducted on diabetes dataset. From the results obtained, it was found that for both 5% training and 10% training SLAD achieved better accuracy of 72.87% compared to the other methods [8].

A method to use sequential variational inference and kalman filtering on diabetes dataset to predict the classification accuracy is proposed [9]. From the output of the method, it was clear that sequential variational inference achieved better accuracy of 80% compared to 76% achieved by kalman filtering.

By combining the advantages of graph and combinatorial method, a clustering ensemble method was developed using Dempster-Shefer evidence theorem [10]. The model was implemented on diabetes dataset and from the experimental results it was identified that the proposed theorem achieved better accuracy of 69.27% compared to other methods.

A method similar to principal component analysis was used to select the important attributes was developed [11]. These attributes are given as an input to the feed forward artificial neural network. The result achieved by the method is measured up with other methods of the feature selection like Tarr's, RUCK's, principal component analysis and t-test. The new model was applied on the diabetes dataset. Testing is done using 20% of data and remaining 80% is used for training. The proposed method achieved good accuracy of 75.22% with less number of attributes.

A selective bayesianclassifier is proposed and is implemented on diabetes dataset using 5-fold cross validation sample [12]. The augmented bayesianclassifier is also implemented on the same dataset. The result of selective bayesianclassifier is compared with naïve bayes and augmented bayesianclassifier. From the result it was clear that selective bayesianclassifier gives better accuracy than the naïve bayes and augmented bayesianclassifier. In addition to this, the selective bayesianclassifier achieves better accuracy of 79.94% through lesser amounts of attributes thus by reducing the size of the dataset.

For inductive concept learning an Evolutionary Concept Learner (ECL) was developed and three different selection mechanisms of ECL: US (US selection operation), weighted US (WUS) and exponentially weighted US (EWUS) were implemented on diabetes dataset [13]. From the result it was found that the average accuracy achieved by EWUS was 77% and better than compared to US and WUS.

A model that makes use of genetic algorithm to select important features is developed in parallel with mapreduce framework [14]. The selected features are produced to k-Nearest Neighbor classifier. The experiment is carried out on diabetes dataset. The accuracy of fitness is calculated using k-Nearest Neighbor. From the result it was seen that parallel genetic algorithm produces better accuracy of 80.51%.

A powerful method is proposed for low dimensional classification and estimation of regression problems [15]. Classification difficulty may be considered as a difficulty of approximating the training set. A multi resolution framework is built based on approximations and organized in the form of a tree. This supports for efficient training. The model is experimented on diabetes dataset and achieves a good accuracy.

An artificial immune recognition system which can notice the existence or nonexistence of disease is developed. The diabetes dataset is run on the machine on an average of 3 runs using 10-fold cross validation sample. The capability of the model is compared with the supplementary methods like IncNEt, Logdisc and Dipol92. The accuracy obtained by the proposed system was 74.1% and was better than the others [16].

A growing-pruning spiking neuron network consisting of 2 stage learning algorithm is developed for handling the problems of pattern classification. The proposed network is consisted of three layers and two stages of learning algorithm and experimented on diabetes dataset [17]. The outcomes are evaluated with batch and online spiking neuron. From the result, it was identified that proposed growing-pruning spiking neural network achieved better accuracy of 71.1%.

Data mining methods like logistic regression and artificial neural networkswith feature selection methods like forward selection and backward elimination are applied on diabetes dataset based on the entropy evaluation method [18]. The experiment was conducted using WEKA. From the result it was identified that the neural network with backward elimination using percentage split achieved an accuracy of 78.90%.

Data mining techniques like logistic regression and artificial neural network are applied on diabetes dataset with feature selection methods like forward selection and backward elimination based on the mean value of the attributes [19]. From the experiment result it was found that an accuracy of 80.46% is achieved by logistic regression when compared to neural network.

Using the threshold value of each attribute an experiment was conducted on diabetes dataset and the performances of the data mining methods like logistic regression and artificial neural networks are evaluated with feature selection methods. From the result it was identified that logistic regression using backward elimination achieved an of accuracy of 82.81% when compared to neural network [20].

Using neural network with back propagation and different data mining techniques like J48, naïve bayes and SVM are applied on diabetes dataset to predict the presence or absence of diabetes in a person. A 5-fold cross validation sample is used to improve the performance of the model. Based on the experimental result conducted an accuracy of 83.11% was achieved by back propagation algorithm [21].

In medical field to exploit the patients information, classification systems are widely used on diabetes dataset. The naïve bayes is applied for classification and for attribute selection genetic algorithm is used [22]. From the experimental results an accuracy of 78.69% is achieved. Popular techniques like deep neural networks and SVM are used to identify the presence or absence of diabetes based on the accuracy of cross validation sample on diabetes dataset. An accuracy of 77.86% was achieved from the said method [23].

Using feature selection for classification of the data has a large number of benefits: decrease in computational difficulty due to decrease in dimensionality [24] and reduction in noise to enhance the classification accuracy [25]. From the literature survey we can identify that many different methods are

applied on the diabetes dataset. Some methods using full set of attributes and some uses the subsets of the attributes. The classification accuracy achieved is not satisfactory and it can be further improved. In this research work we try to improve the accuracy by using the data mining technique like random forest withfeature selection methods like forward selection and backward eliminationusingpercentage split as test option. In the next section we study the proposed framework of the research carried out.

3. PROPOSED FRAMEWORK

The proposed framework for evaluating the PIMAIndian Diabetes dataset with feature selection methods using R is shown in Figure 1. The process of evaluation is as follows:

a. The first step is the selection of the diabetes dataset.
b. For any missing values in the dataset, pre-processing is done. Since the dataset considered have no missing values, so no pre-processing is required. The dataset is taken in its original form.
c. We find the entropy value of each attribute of the dataset using (1)

$$\boldsymbol{Info(D)= \sum_{i=1}^{m-1} p_i \log_2(p_i)} \quad (1)$$

where D is the attribute, i is the attribute index, p_i is the probability that an attribute in D belongs to a class and m is the total count of attributes.

d. Not all the features in the dataset are important in prediction. Based on the entropy value of each attribute, apply feature selection methods like forward selection and backward elimination to obtain the different subsets of features.
e. For each subset of feature we evaluate the performance of random forest using percentage split as test option.
f. Finally the subset of features which achieves better accuracy are considers as very important attributes.

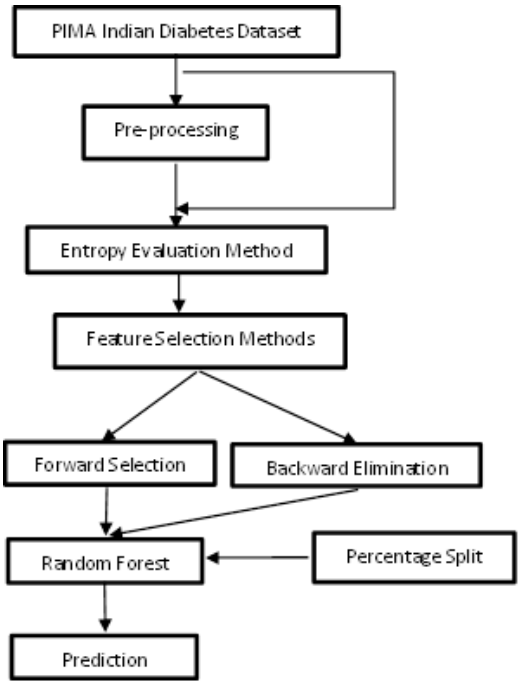

Figure 1. Framework for the proposed work based on random forest using percentage split

4. RESULTS AND DISCUSSION

In this research work we evaluate the performance of the data mining method like random forest with feature selection methods using percentage split as test option on diabetes dataset using R. The diabetes dataset considered in the research work consists of 768 instances and 9 attributes. Table 1 gives the list of attribute in dataset and their meaning. Table 2 and Table 3 gives the different subsets of attributes obtained after applying the feature selection methods and the best accuracy achieved by random forest.

Table 1. Diabetes dataset attributes and meaning

Attribute	Meaning
preg	Number of times pregnant
plas	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
pres	Diastolic blood pressure (mm Hg)
skin	Triceps skin fold thickness (mm)
insu	2-Hour serum insulin (mu U/ml)
mass	Body mass index (weight in kg/(height in m)^2)
pedi	Diabetes pedigree function
age	Age (years)
class	Class variable (0 or 1)

Table 2. Different subsets obtained and the best accuracy achieved after applying forward selection based on entropy value

Subset No.	Subset of Attributes	No. of Attributes	Classification Accuracy
1	pres, class	2	57.4%
2	pres, pedi, class	3	58.5%
3	preg, pres, pedi, class	4	62%
4	preg, pres, skin, pedi, class	5	63.6%
5	preg, pres, skin, insu, pedi, class	6	68%
6	preg, pres, skin, insu, pedi, age, class	7	73.2%
7	preg, pres, skin, insu, mass, pedi, age, class	8	75.5%
8	Full set of attributes	9	83.8%

Table 3. Different subsets obtained and the best accuracy achieved after applying backward elimination based on entropy value

Subset No.	Subset of Attributes	No. of Attributes	Classification Accuracy
1	preg, plas, skin, insu, mass, pedi, age, class	8	84.1%
2	preg, plas, skin, insu, mass, age, class	7	83.1%
3	plas, skin, insu, mass, age, class	6	81.4%
4	plas, insu, mass, age, class	5	81.7%
5	plas, mass, age, class	4	80.8%
6	plas, mass, class	3	74.6%
7	plas, class	2	71.3%

From the classification accuracy obtained for different set of attributes as shown in Table 2 and Table 3, we identified that:

a. The accuracy achieved for full set of attributes is 83.8%.
b. The best classification achieved is by Subset No. 1 of Table 3 is 84.1%, which is better than the accuracy achieved for full set of attributes and other methods applied on diabetes dataset. The comparison between the accuracy achieved by the proposed method and some of the existing methods is shown in Figure 2.

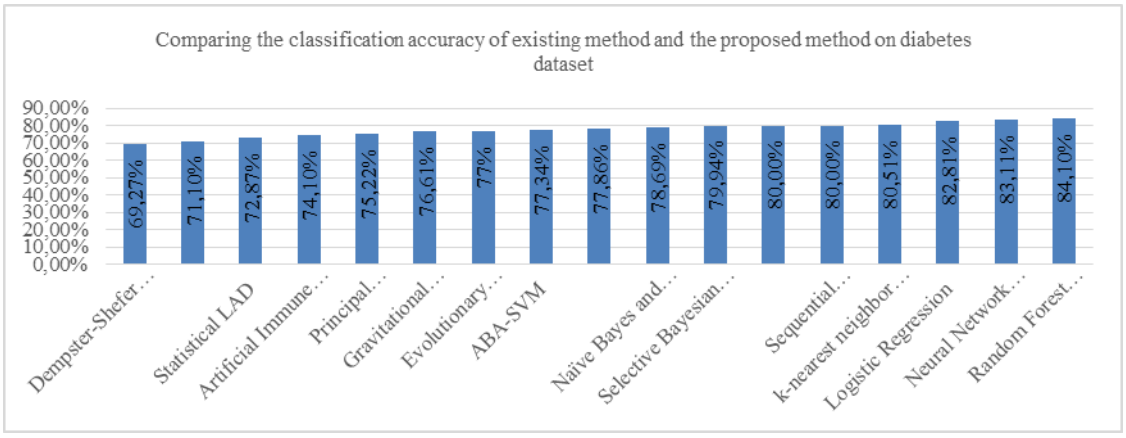

Figure 2. Comparison of classification accuracy of the existing method with the proposed method

5. CONCLUSION

In this research work the data mining method like random forest with feature selection methods like forward selection and backward elimination is applied on diabetes dataset using percentage split as test option. The experiment is implemented in R studio using R programming languages. From the experimental results we can observe the following:

a. The classification accuracy achieved for full set of attributes is 83.8%.
b. The classification accuracy achieved by the proposed method is 84.1% with 7 attributes namely preg, plas, skin, insu, mass, pedi, and age.
c. The accuracy achieved is better than any of the existing methods on diabetes dataset and the accuracy achieved by using full set of attributes as shown in Figure 2.

Further researchers can compare the Classification Accuracy with other techniques like SVM, NN, and NN with K fold, deep learning and may achieve better Classification Accuracy.

ACKNOWLEDGEMENTS

I would like express my deep gratitude to the HoD and Staff of Computer Science and Engineering department of CHRIST Deemed to be University, Bangalore for supporting me in doing this research work.

REFERENCES

[1] S. K. Wasan, *et al*., "The Impact of Data Mining Techniques on Medical Diagnostics," *Data Science Journal*, vol. 5, pp. 119-126, 2006.
[2] T. K. Ho, "The Random Subspace Method for Constructing Decision Forests," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 832-844, 1998.
[3] W. G. Touw, *et al*., "Data Mining in the Life Sciences with Random Forest: a Walk in the park or lost in Jungle?" *Briefings in Bioinformatics*, vol. 14, pp. 315-326, 2012.
[4] R. Abraham, *et al*., "Effective Discretization and Hybrid Feature Selection Using Naïve Bayesian Classifier for Medical Data Mining," *International Journal of Computational Intelligence Research*, vol. 5, pp. 116-129, 2009.
[5] M. L. Raymer, *et al*., "Knowledge Discovery in Medical and Biological Datasets Using a Hybrid Bayes Classifier/ Evolutionary Algorithm," *IEEE Transactions on Systems, Man and Cybernatics*, vol. 33, pp. 802-813, 2003.
[6] M. B. Dowlatshahi and M. Rezaeian, "Training Spiking Neurons with Gravitational Search Algorithm for Data Classification," *IEEE 1st International Conference on Swarm Intelligence and Evolutionary Computation*, pp. 53-58, 2016.
[7] E. Tuba, *et al*., "Adjusted Bat Algorithm for Tuning of Support Vector Machine Parameters," *IEEE Congress on Evolutionary Computation*, pp. 2225-2232, 2016.
[8] R. Bruni and G. Bianchi, "Effective Classification Using a Small Training Set Based on Discretization and Statistical Analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, pp. 2349-2361, 2015.
[9] P. Sykacek and S. Roberts, "Adaptive Classification by Variational Kalman Filtering," *Advances in Neural Information Processing Systems (NIPS)*, pp. 737-744, 2002.
[10] F. J. Li, *et al*., "Multigranulation Information Fusion: A Dempster-Shafer Evidence Theory Based Clustering Ensemble Method," *Proceedings of IEEE International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 1, pp. 58-63, 2015.
[11] S. J. Perantonis and V. Virvilis, "Input Feature Extraction for Multilayered Perceptrons Using Supervised Principal Component Analysis," *Neural Processing Letters*, vol. 10, pp. 243-252, 1999.
[12] C. A. Ratanamahatana and D. Gunopulos, "Feature Selection for the Naïve Bayesian Classifier Using Decision Trees," *Applied Artificial Intelligence (AAI)*, vol. 17, pp. 475-487, 2003.
[13] F. Divina and E. Marchiori, "Knowledge-Based Evolutionary Search for Inductive Concept Learning," *Knowledge Incorporation in Evolutionary Computation, Springer*, vol. 167, pp. 237-253, 2005.
[14] G. T. Hilda and R. R. Rajalaxmi, "Effective Feature Selection for Supervised Learning Using Genetic Algorithm," *IEEE 2nd International Conference on Electronics and Communication Systems (ICECS 2015)*, pp. 909-914, 2015.
[15] I. Blayvas and R. Kimmel, "Machine Learning via Multiresolution Approximations," *IEICE Transaction on Information System*, vol. E86-D, pp. 1172-1180, 2003.
[16] A. Watkins, *et al*., "Artificial Immune Recognition System (AIRS): an Immune Inspired Supervised Learning Algorithm," *Genetic Programming and Evolvable Machines*, vol. 5, pp. 291-317, 2004.
[17] S. Dora, *et al*., "A Two Stage Learning Algorithm for a Growing-Pruning Spiking Neural Network for Pattern Classification Problems," *International Joint Conference on Neural Networks (IJCNN)*, pp. 1-7, 2015.
[18] S. Raghavendra and M. Indiramma, "Performance Evaluation of Logistic Regression and Artificial Neural Network Model with Feature Selection Methods Using Cross Validation Sample and Percentage Split on Medical Datasets," *International Conference on Emerging Research in Computing, Information, Communication and Applications*, vol. 2, 2014.
[19] S. Raghavendra and M. Indiramma, "Classification and Prediction Model using Hybrid Technique for Medical Datasets," *International Journal of Computer Applications*, vol. 127, pp. 20-15, 2015.

[20] S. Raghavendra and M. Indiramma, "Hybrid Data Mining Model for the Classification and Prediction of Medical Datasets," *International Journal of Knowledge Engineering and Soft Data Paradigms*, vol. 5, pp. 262- 284, 2017.
[21] F. G. Woldemichael and S. Menaria, "Prediction of Diabetes Using Data Mining Techniques," *2nd International Conference on Trends In Electronics and Informatics*, pp. 414-418, 2018.
[22] D. K. Choubey, *et al*., "Classification of Pima Indian Diabetes Dataset Using Naïve Bayes with Genetic Algorithm as an Attribute Selection," *Communication and Computing Systems, Taylor & Francis Group*, pp. 451-455, 2017.
[23] S. Wei, *et al*., "A Comprehensive Exploration to the Machine Learning Technique for Diabetes Dataset," *IEEE 4th World Forum on Internet of Things*, pp. 291-295, 2018.
[24] R. Abraham, *et al*., "Effective Discretization and Hybrid Feature Selection Using Naïve Bayesian Classifier for Medical Data Mining," *International Journal of Computational Intelligence Research*, vol. 5, pp. 116-129, 2009.
[25] Q. Cheng, *et al*., "Logistic Regression for Feature Selection and Soft Classification of Remote Sensing Data," *IEEE Geoscience and Remote Sensing Letters*, vol. 3, pp. 491-494, 2006.

BIOGRAPHIES OF AUTHORS

Dr. Raghavendra S is currently working as Associate Professor in the Department of Computer Science and Engineering at CHRIST Deemed to be University, Bangalore. He completed his Ph.D. degree in Computer Science and Engineering from VTU, Belgaum, India in 2017 and has 14 years of teaching experience. His interests include Data Mining and Big data.

Santosh Kumar J is currently working as Associate Professor in the Department of Computer Science and Engineering at K.S.School of Engineering and Management, Bangalore. He is pursuing Ph.D. in VTU, Belgaum, India. He has 10 years of teaching and 3 years of industry experience. He is specialized in Big data streaming analysis. His reserach topics includesBig data with machine learning.